Learning joint shape and appearance representations with metamorphic auto-encoders

Alexandre Bône, Paul Vernhet, Olivier Colliot, Stanley Durrleman

Learning joint shape and appearance representations with metamorphic auto-encoders

Alexandre Bône[†], Paul Vernhet[†], Olivier Colliot, and Stanley Durrleman

ARAMIS Lab, ICM, Inserm U1127, CNRS UMR 7225, Sorbonne University, Inria, Paris, France
{firstname.lastname}@icm.institute.org

Abstract. Transformation-based methods for shape analysis offer a consistent framework to model the geometrical content of images. Most often relying on diffeomorphic transforms, they lack however the ability to properly handle texture and differing topological content. Conversely, modern deep learning methods offer a very efficient way to analyze image textures. Building on the theory of metamorphoses, which models images as combined intensity-domain and spatial-domain transforms of a prototype, we introduce the "metamorphic" auto-encoding architecture. This class of neural networks is interpreted as a Bayesian generative and hierarchical model, allowing the joint estimation of the network parameters, a representative prototype of the training images, as well as the relative importance between the geometrical and texture contents.
We give arguments for the practical relevance of the learned prototype and Euclidean latent-space metric, achieved thanks to an explicit normalization layer. Finally, the ability of the proposed architecture to learn joint and relevant shape and appearance representations from image collections is illustrated on BraTs 2018 datasets, showing in particular an encouraging step towards personalized numerical simulation of tumors with data-driven models.

Keywords: Numerical brain atlas · Shape analysis · Metamorphosis

1 Introduction

The shape analysis literature offers a number of tools to perform statistical analysis tasks on geometrical objects. At the core of the founding works [12, 14] lies the idea to quantify the differences between two shapes thanks to large parametric classes of deformations that warp one into the other: once the optimal transformation found, its norm provides a proxy distance metric. Diffeomorphic transformations are for instance widely used for medical image analysis [31], with applications to image registration or atlas building [1,25,32]. However, these transformations are purely geometrical, and cannot account for potential texture (or "appearance") variability in the considered images. In particular, images with differing topological contents cannot be diffeomorphically warped one into the

[†]Equal contributions.

other. This limitation gave birth to the early theoretical work [29] where the proposed "metamorphoses" jointly transform the geometry and the intensity of an image. If some authors built on this idea for brain tumor monitoring [23], sub-cortical brain segmentation [24] or learning generalized principal component analysis models [3], the literature is fairly limited.

Conversely, if modern deep learning architectures are commonly believed to aggregate information up until deep filters able recognize entire shapes [19, 21], some recent works [6, 8, 13] question this so-called shape hypothesis, suggesting that texture (or "appearance") features carry more weight. Beyond the local spatial invariances achieved by max-pooling layers (with receptive fields typically of the order of a few pixels), deep learning methods are agnostic of the data nature: in other words, potential powerful prior knowledge is not taken into account. This agnostic approach presents two disadvantages: the interpretability of the network is often limited, and huge amounts of data might be needed for the network to re-discover already-known data properties.

Most data augmentation techniques are attempts to alleviate this second point, by artificially increasing the data set size according to priors such as invariance to small intensity-domain or affine spatial-domain transformations [9,20,27]. At the cost of a longer training time, the network learns the implicitly-encoded invariance properties. Instead of implicitly suggesting invariances or manipulating data set biases with data augmentation, learning from small data sets and enhanced interpretability can be achieved by designing adapted architectures that explicitly enforce priors [15, 28]. Some recent attempts to fill the gap between classical model-based shape analysis tools and data-driven deep learning methods managed to combine the learning flexibility of the former with the theoretical guarantees and interpretability of the latter. Building on the variational auto-encoding architecture introduced in [16], the approaches described in [11, 18] learn probabilistic diffeomorphic registration models, which cannot handle varying topology in data, and do not allow the construction of an atlas, i. e. a reference image estimated by groupwise registrations among a training image data set. [7] learns atlas models but is bound to the same topological limitations as the previous ones. In [26], the authors go beyond pure spatial-warping layers and further introduce joint spatial and intensity transformations in auto-encoding networks. This first attempt, still only of its kind to the best of our knowledge, relies on an appearance prediction decoder followed by an ad-hoc warping module specifically developed. Once learned, the latent representations can be manipulated to complete disentangled shape or appearance reconstruction and interpolation. This work is the closest to metamorphosis one, however the regularity of the deformation field is controlled by a loss penalty and thus not ensured by design of the network, furthermore no links are made with the rich theoretical background of metamorphosis.

In this paper, we introduce the metamorphic auto-encoders (MAEs) which relies on the prior that training images can be seen as "metamorphic" transformations of a prototype. We show that thanks to this assumption, estimating MAE architectures is equivalent to learning disentangled shape and appearance

low-dimensional representations from imaging data sets in an unsupervised fashion. Thanks to an isometry-enforcing layer inspired by the underlying theory, the learned representations are embedded in a relevant metric space where readily-available Euclidean operations potentially allow to perform image manipulation. This introduced class of neural networks is interpreted as a Bayesian generative and hierarchical model, allowing the joint estimation of the network parameters, a representative prototype of the training images, as well as the relative importance between the geometrical and texture contents. This work is therefore a point of convergence between Bayesian generative statistics, metamorphoses-based shape analysis and variational auto-encoding methods.

Section 2 details the chosen transformation model, from its geometrical interpretation to its practical implementation. Section 2.3 presents the proposed MAE architecture, its Bayesian probabilistic interpretation, and the chosen optimization scheme. Section 3 shows the ability of MAE to learn relevant disentangled shape and appearance representations of imaging data sets and illustrates the potential of allowed post-processing image manipulation applications. Section 4 concludes.

2 Metamorphic transformation model

We first detail briefly our theoretical metamorphic transformation model, which builds on different fields of the shape analysis literature [2, 29, 31]. Illustrated by Figure 1, its discrete counterpart is then derived to prepare the integration of metamorphoses into neural network architectures. See *Supplementary materials* for a self-contained minimal presentation of the metamorphosis framework.

2.1 Continous theory

Let $\Omega \subset \mathbb{R}^d$ with $d \in \{2, 3\}$ be a spatial domain, on which is defined the image $I_0 : \Omega \to \mathbb{R}$. This image can be deformed into $\phi_1 \star I_0 = I_0 \circ \phi_1$ by any diffeomorphism ϕ_1 of Ω. Similarly to [2], we choose to construct diffeomorphisms by following the flow of static and smooth velocity vector fields $v \in V \subset C_0^\infty(\Omega, \mathbb{R}^d)$ for a unit time period $t \in [0, 1]$, i.e. by integrating:

$$\partial_t \phi_t = v \circ \phi_t \quad \text{from the identity} \quad \phi_0 = \mathrm{Id}_\Omega. \tag{1}$$

We denote $\Phi : v \to \phi_1$ the mapping which associates the obtained diffeomorphism from the vector field v. Abusing slightly of notations, for any velocity field v and intensity increment $\delta \in D \subset C_0^\infty(\Omega, \mathbb{R})$, the action of the metamorphosis operator $\Phi(v, \delta)$ on images is given by:

$$\Phi(v, \delta) \star I_0 = \Phi(v) \star (I_0 + \delta) = (I_0 + \delta) \circ \Phi(v). \tag{2}$$

A new metamorphic distance $d_{I_0} \geq 0$ on images can be defined as:

$$d_{I_0}(I, I')^2 = \|v' - v\|_V^2 + \|\delta' - \delta\|_D^2 \text{ where } (v, \delta) = \Phi^{-1}(I) \text{ and } (v', \delta') = \Phi^{-1}(I'). \tag{3}$$

whose norms can be computed by further considering than v and δ live in reproducing kernel Hilbert spaces with kernels K_V and K_D respectively.

2.2 Practical discrete case

In practice, the physical domain Ω is discretized into a regular grid g, and the time segment $[0,1]$ into 2^T time-points, with $T \in \mathbb{N}^*$. We choose K_V and K_D as simple radial Gaussian kernels of respective radiuses $\rho_V > 0$ and $\rho_D > 0$. The g-discretized fields $\underline{v}^*$, $\underline{\delta}^*$ and atlas $\underline{I}_0$ are fed as inputs to the discrete metamorphic transformation module. As illustrated by the right-side of Figure 1, the discrete velocity field $\underline{v}$ and intensity increment $\underline{\delta}$ are first computed according to the filtering formulae $\underline{v} = \underline{K}_V \cdot \underline{v}^*$ and $\underline{\delta} = \underline{K}_D \cdot \underline{\delta}^*$ where for any grid index k_0:

$$\big[\underline{K}_V\big]_{k_0} = \sum_k \exp\left(\frac{-\|g_k - g_{k_0}\|^2_{\ell^2}}{\rho_V^2}\right) \text{ and } \big[\underline{K}_D\big]_{k_0} = \sum_k \exp\left(\frac{-\|g_k - g_{k_0}\|^2_{\ell^2}}{\rho_D^2}\right) \tag{4}$$

which corresponds to the discrete version of the reproducing Hilbert norm, which writes $\|\underline{v}\|_V = (\underline{v}^*)^\top \cdot \underline{K}_V \cdot \underline{v}^*$ (and similarly on D). As originally described in [2], the integration along the streamlines of v defined by Equation 1 is discretely carried out with the scaling-and-squaring algorithm which consists in applying T times:

$$\underline{x}_{t+1} = \underline{x}_t + \mathcal{I}(\underline{x}_t - g,\, \underline{x}_t) \qquad \text{from} \quad \underline{x}_0 = g + \underline{v} \,/\, 2^T \tag{5}$$

where $\mathcal{I}(\underline{x}_t - g,\, \underline{x}_t)$ simply denotes the interpolation of the displacement field $\underline{x}_k - g$ at the physical locations $\underline{x}_k$. The metamorphosis of I_0 (i.e. Equation 2) is finally approximated as:

$$\Phi(v, \delta) \star I_0 \;\approx\; \Phi(\underline{v}, \underline{\delta}) \star \underline{I}_0 \;=\; \mathcal{I}(\underline{I}_0 + \underline{\delta},\, \underline{x}_T) \tag{6}$$

where $\mathcal{I}(\underline{I}_0 + \underline{\delta},\, \underline{x}_T)$ here denotes the interpolation[†] of the intensity values $\underline{I}_0 + \underline{\delta}$ at locations $\underline{x}_T$.

2.3 Variational formulation for generative modeling

Our statistical model is based on the variational framework [16, 17], and consists in assuming that the observed images $(I_i)_{i=1}^n$ are hierarchically distributed according to

$$I_i \overset{\text{iid}}{\sim} \mathcal{N}\Big\{\Phi\big[S_\sigma(s_i),\, A_\alpha(a_i)\big] \star I_0;\; \epsilon^2\Big\} \text{ with } s_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \lambda_s^2) \text{ and } a_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \lambda_a^2) \tag{7}$$

where I_0 is the learned atlas, S_σ (respectively A_α) is a σ-parametric (respectively α-parametric) neural network mapping, that associates the velocity field v (respectively the intensity increment δ) to any code $s \in \mathbb{R}^p$ (respectively $a \in \mathbb{R}^q$).

Figure 1 details the architecture of the metamorphic auto-encoder. The first important observation is that we structurally[‡] impose the metric equivalence

[†] we used bilinear interpolation scheme

[‡] via an explicit normalization layer

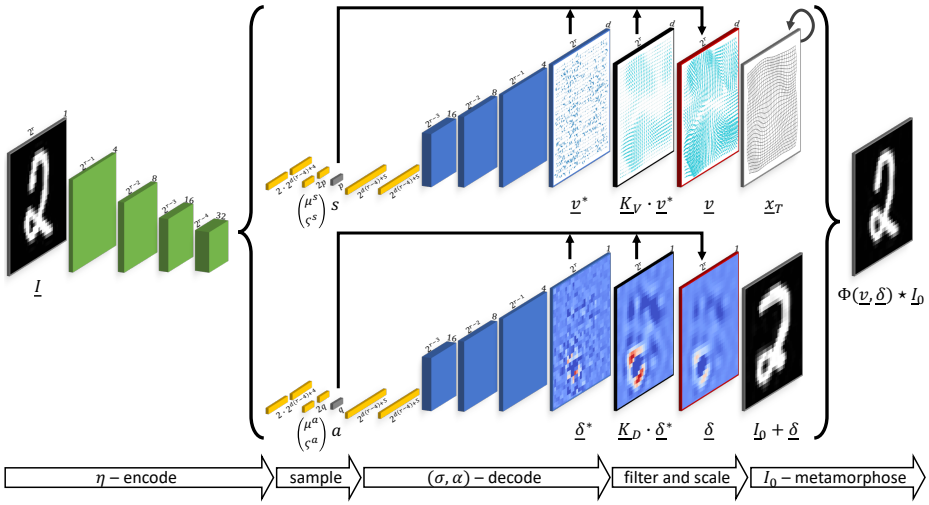


Fig. 1: Architecture of the metamorphic auto-encoder. The input image $\underline{I}$ is encoded by four convolution layers (in green), followed by two parallel pairs of fully-connected layers (in yellow). The encoder outputs are interpreted as characterizing two normal distributions, from which are sampled the latent codes $s \in \mathbb{R}^p$ and $a \in \mathbb{R}^q$. Two parallel decoders successively composed of two fully connected and four deconvolution layers map those latent shape and appearance representations to the velocity field $\underline{v}^*$ and intensity increment $\underline{\delta}^*$ duals. After filtering by the operators $\underline{K}_V$ and $\underline{K}_D$, the obtained vectors are explicitly scaled, enforcing the equality of their Hilbert norm with the Euclidean norm of the corresponding codes s and a (see Equation 4). The resulting velocity field $\underline{v}$ (see Equation 5) and intensity increment $\underline{\delta}$ (see Equation 6) are finally combined to metamorphose the prototype image parameter $\underline{I}_0$.

between the metamorphic distance d_{I_0} defined by Equation 3 and the induced latent norm $d_0(z, z') = \|z - z'\|_{\ell^2}$: the mappings are therefore isometric, i.e. that $\|v\|_V = \|s\|_{\ell^2}$ and $\|\delta\|_D = \|a\|_{\ell^2}$.

Furthermore, we chose tanh activation functions after convolutions (at the exception of the last encoding one) and deconvolutions, and that all decoding layers are chosen without bias : these two last hypotheses ensure the infinite differentiability of the mappings $\underline{S}_\sigma$, $\underline{A}_\alpha$ and $\underline{E}_\eta$.

Lastly, we chose to encode the euclidean difference $I_i - I_0$ as input in our network, which associated with null bias and tanh activations imposes that the null latent-space vector $z = 0$ is mapped to I_0; the prior distribution defined on the random effects $(z_i)_i$ therefore defines I_0 as a statistical average of the observations $(I_i)_i$. In other words, estimating the model parameters θ and I_0 can legitimately be interpreted as computing a Fréchet average of the training data set [25].

3 Experiments

The ability to disentangle shape and appearance becomes necessary if we want to manipulate data where abnormalities, such as tumors, appears at a visible scale. We applied our unsupervised framework to the task of reconstructing brains with tumors, in the spirit of personalized numerical modeling, on BraTs 2018 dataset [4, 5, 22]. Images have been obtained by selecting an axial section of T1 contrast enhanced 3D brain volumes, pre-processed with standard skull-removing and affine alignment pipelines.

Out first series of experiments compares behavior of metamorphic model to its known diffeomorphic equivalent [7] : Fig 2 depicts how the flow of transformation for each model works. A comparison with the diffeomorphic equivalent model to ours shows clearly the necessity to disentangle : without the δ module, atlas learned has no proper geometrical features, in addition to a poor reconstruction of brain. Indeed the tumors are obtained through inclusion of saturated points on the learned atlas when diffeomorphism only are used, and even then reconstruction requires high deformations. On the contrary, metamorphisms naturally handle the separation of information, with smoother velocity fields, clear localization of the tumor on the intensity map δ as well as a visually sharp atlas representing a brain without tumors. As a sanity check, simple classification of tumor grade (high vs low) was performed taking the various latent representations learned as features and summarized on table 1.

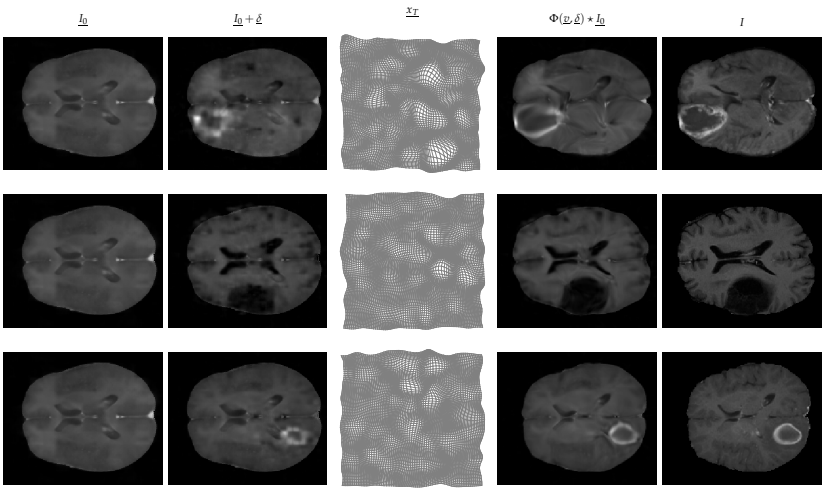
(%)	Full latent space	Shape latent space	Appearance latent space
Metamorphic	64.0 ± 13.0	64.0 ± 15.0	67 ± 12.0
Diffeomorphic	57.0 ± 16.0	-	-

Table 1: Average quadratic discriminant analysis (QDA) balanced accuracy classification scores (stratified 10-fold method, chance level 50%).

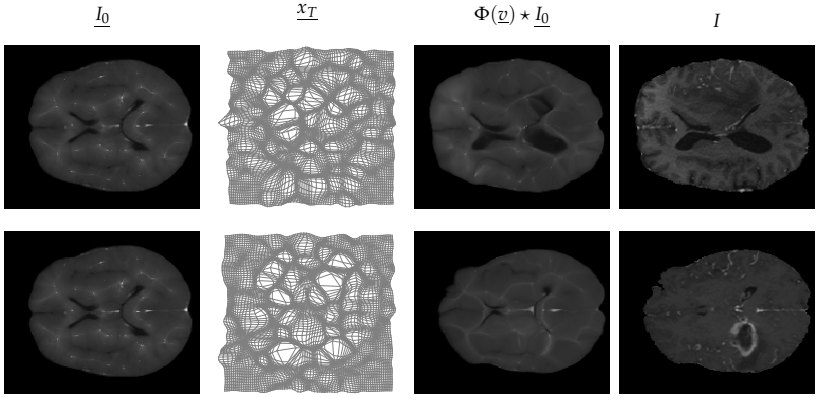
For a more quantitative understanding, we also applied the Chan Vese level-set segmentation algorithm [10] on δ maps to quantitatively assess the quality of the learned atlas : relatively high Dice scores, up to 0.7 can be attained by segmenting this map and transporting, through v, the obtained mask. Fig 3 shows examples of obtained masks. This experiment suggests than intrinsic information of the tumor was mainly grasped by the δ (for the appearance part) and v (for the geometric localization) components of our model, leaving the atlas $\mathcal{I}_0$ to be understood as a control-like Frechet-mean.

4 Conclusion

Our method may be understood at the crossroad of simulation numerical models, which are based on sound mathematical priors (metamorphosis decomposition



(a) Metamorphic transformation module.



(b) Equivalent diffeomorphic model

Fig. 2: Comparison of diffeomorphic and metamorphic models

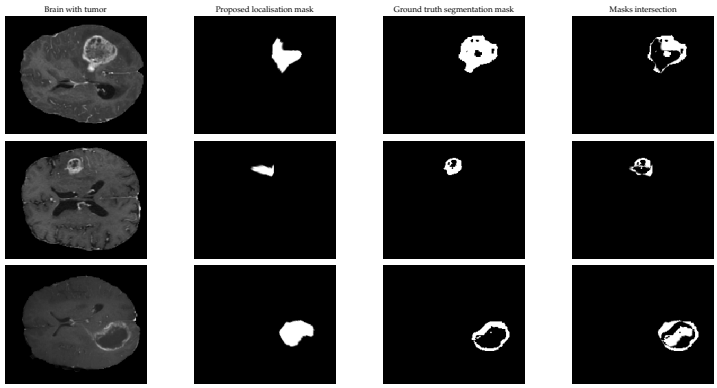


Fig. 3: Unsupervised localisation of tumor using Chan-Vese segmentation algorithm [10] on appearance maps δ (on this figure, $DICE \geq 0.60$), from left to right: input brain, computed segmentation mask, ground-truth tumor mask, intersection of masks

of brain variability), and data-driven statistics which may soon enable efficient generative models to emerge, then applicable to personalize individual evolution of pathology.

Untangling the shape and appearance variabilities is of key interest in computational anatomy to better interpret and apprehend the total variability of a collections of organs or anatomical shapes: better disease markers for medical images could for instance be identified, and in the case of tumors improved growth models could be developed.

Qualitative results were presented aiming towards well-behaved models evidence, in particular the segmentation of δ-map which led to very coherent localization of tumors without complex treatment nor specific dataset filtering. To the best of our knowledge, no publicly available code was developed for formal metamorphosis framework, which therefore didn't allow for a direct comparison with our deep-learning method, though a natural advantage is its sampling efficiency and thus potential use in numerical simulation scenarios or for data augmentation.

Lastly, our network was composed in a very simple fashion, and increasing the architecture complexity, in particular through powerful U-net structures [30] will naturally allow to reach stronger reconstruction accuracies and better performances.

References

1. Allassonnière, S., Durrleman, S., Kuhn, E.: Bayesian mixed effect atlas estimation with a diffeomorphic deformation model. SIAM Journal on Imaging Science **8**, 1367–1395 (2015)
2. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 924–931. Springer (2006)
3. Ashburner, J., Brudfors, M., Bronik, K., Balbastre, Y.: An algorithm for learning shape and appearance models without annotations. arXiv preprint arXiv:1807.10731 (2018)
4. Bakas, S., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific Data **4** (09 2017). https://doi.org/10.1038/sdata.2017.117
5. Bakas, S., et al.: Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv e-prints arXiv:1811.02629 (Nov 2018)
6. Ballester, P., Araujo, R.M.: On the performance of googlenet and alexnet applied to sketches. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
7. Bône, A., Louis, M., Colliot, O., Durrleman, S.: Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders (2019)
8. Brendel, W., Bethge, M.: Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint arXiv:1904.00760 (2019)
9. Cireşan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. arXiv preprint arXiv:1202.2745 (2012)
10. Cohen, R.: The Chan-Vese Algorithm. arXiv e-prints arXiv:1107.2782 (Jul 2011)
11. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning for fast probabilistic diffeomorphic registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 729–738. Springer (2018)
12. D'Arcy Wentworth, T.: On growth and form. Abridged ed.(Tyler Bonner, John ed.) Cambridge University Press. (1917)
13. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
14. Grenander, U.: General pattern theory-A mathematical study of regular structures. No. BOOK, Clarendon Press (1993)
15. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. stat **1050**, 10 (2014)
17. Kingma, D.P., Welling, M.: An Introduction to Variational Autoencoders. arXiv e-prints arXiv:1906.02691 (Jun 2019)
18. Krebs, J., Delingette, H., Mailhé, B., Ayache, N., Mansi, T.: Learning a probabilistic model for diffeomorphic registration. IEEE Transactions on Medical Imaging (2019)
19. Kriegeskorte, N.: Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual review of vision science **1**, 417–446 (2015)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

21. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436 (2015)
22. Menze, B., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **99** (12 2014). https://doi.org/10.1109/TMI.2014.2377694
23. Niethammer, M., Hart, G.L., Pace, D.F., Vespa, P.M., Irimia, A., Van Horn, J.D., Aylward, S.R.: Geometric metamorphosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 639–646. Springer (2011)
24. Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M.: A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage **56**(3), 907–922 (2011)
25. Pennec, X.: Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. Journal of Mathematical Imaging and Vision **25**(1), 127–154 (2006)
26. Shu, Z., Sahasrabudhe, M., Alp Guler, R., Samaras, D., Paragios, N., Kokkinos, I.: Deforming autoencoders: Unsupervised disentangling of shape and appearance. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 650–665 (2018)
27. Simard, P.Y., Steinkraus, D., Platt, J.C., et al.: Best practices for convolutional neural networks applied to visual document analysis. In: Icdar. vol. 3 (2003)
28. Skafte Detlefsen, N., Freifeld, O., Hauberg, S.: Deep diffeomorphic transformer networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4403–4412 (2018)
29. Trouvé, A., Younes, L.: Metamorphoses through lie group action. Foundations of Computational Mathematics **5**(2), 173–198 (2005)
30. Tudosiu, P.D., Varsavsky, T., Shaw, R., Graham, M., Nachev, P., Ourselin, S., Sudre, C.H., Cardoso, M.J.: Neuromorphologicaly-preserving Volumetric data encoding using VQ-VAE. arXiv e-prints arXiv:2002.05692 (Feb 2020)
31. Younes, L.: Shapes and Diffeomorphisms. Applied Mathematical Sciences, Springer Berlin Heidelberg (2010), https://books.google.fr/books?id=SdTBtMGgeAUC
32. Zhang, M., Singh, N., Fletcher, P.T.: Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In: IPMI. vol. 23, pp. 37–48 (2013)